Aman Patel

October 26, 2018

UROC @ SICE

Week 4: Extracting Subcategories and Editing URLs

This week I was tasked with extracting the subcategories from a category of a Wikipedia article. To do this, I was required to edit the original URL by changing the article name and the properties that would be extracted. To begin, I followed the same process as last week to extract the list of categories using the following URL:

<https://en.wikipedia.org/w/api.php?format=json&action=query&prop=categories&titles=Python_(programming_language)&cllimit=100>

I then selected the category Programming languages and saved the name of the category into a string. I replaced the space found in the string with an underscore to allow it to be processed as a URL and replaced the article name in the original URL with this string. I then duplicated this URL and edited one copy to extract the list of subcategories. The other copy was kept the same to extract the list of categories of the category selected. Once the URLs were finalized, I followed the same process to extract, save, and print the lists from each URL. The final URLs are shown below:

Categories of selected category: <https://en.wikipedia.org/w/api.php?format=json&action=query&prop=categories&titles=Category:Programming_languages&cllimit=100>

Subcategories of selected category: <https://en.wikipedia.org/w/api.php?format=json&action=query&list=categorymembers&cmtitle=Category:Programming_languages&cmtype=subcat&cmlimit=500>